PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-325917
(43)Date of publication of application : 16.12.1997

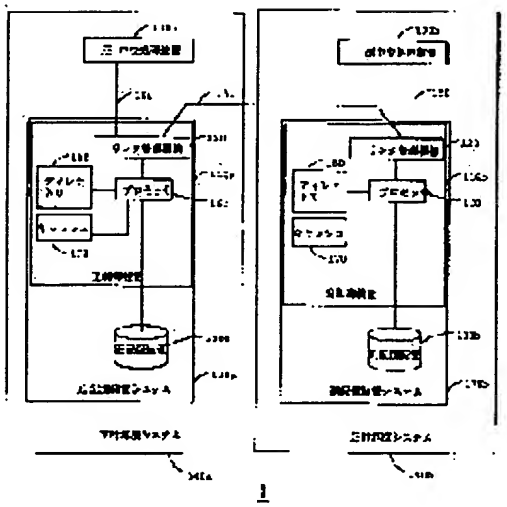
(51)Int.CI. G06F 12/16
G06F 3/06
G06F 3/06

(21)Application number : 08-145562
(22)Date of filing : 07.06.1996

(71)Applicant : HITACHI LTD
(72)Inventor : YAMAKAMI KENJI
NAKAMURA KATSUNORI
YAMAMOTO AKIRA

(54) COMPUTER SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a computer system in which whether data are last or not can be recognized by a subordinate storage device when a regular controller can not be used because of any fault or disaster in the case of applying an asynchronous copy system.
SOLUTION: In accordance with a write request from a central processing unit 100, a regular controller 110a transfers only the position information of write object data to a subordinate controller 110b. At the subordinate controller 110b, the position information is held until data are actually written. When a regular storage device 130a can not be accessed because of any fault or disaster, the position information is read out of the subordinate controller 110b so that the presence/absence of erased data, its position or the time can be discriminated.



1

LEGAL STATUS
[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

JP09-325917

* NOTICES *
JPO and NCIPI are not responsible for any
damages caused by the use of this translation.
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] In the computing system which consists of a forward storage system which connects with a central processing unit and consists of storage and a control unit, and a remote substorage system linked to the forward storage system the control unit, i.e., the forward control unit, under said forward storage system When a light demand is received from said central processing unit, after transmitting the positional information to said substore, without transmitting light data, light completion is reported to said central processing unit. Light processing to said substorage is performed at a suitable next stage., the control unit, i.e., the secondary controller, under said substorage system When said positional information transmitted from said forward control unit is held, light processing of being behind actual is performed, said positional information is canceled and said forward control unit becomes use impossible according to disaster etc. The computing system characterized by making existence of the data which disappeared detectable by said positional information stored in said secondary controller.
[Claim 2] In a computing system according to claim 1 said forward control unit To said positional information, in addition, by transmitting the time of day which had the light demand from said central processing unit to said secondary controller When said secondary controller is made to memorize and said forward control unit becomes use impossible according to disaster etc., the renewal sequence of data of said central processing unit The computing system characterized by determining the backup file which should be recovered and performing data recovery from the time of day stored in said secondary controller.
[Claim 3] It is the computing system characterized by said secondary controller reporting error termination to said central processing unit when the field corresponding to said positional information has access from said central processing unit to said secondary controller in a computing system according to claim 1.
[Claim 4] In the computing system which consists of a forward storage system which connects with a central processing unit and consists of storage and a control unit, and a remote substorage system linked to said forward storage system the control unit under said forward storage system, i.e., a forward control unit, and the control unit, i.e., the secondary controller, under said substorage system Two conditions, i.e., a duplex writing coincidence condition, and a duplex writing inequality condition are managed as a condition [writing / duplex], respectively. Said forward control unit If the forward storage for a light is in a duplex writing coincidence condition when a light demand is received from said central processing unit Make said each duplex writing condition of said forward storage and said substorage change in the duplex writing inequality condition, and light data report light completion to said substorage to said central processing

unit, without transmitting. Light processing to said substorage is performed at a suitable next stage. Further When a duplex writing coincidence condition is made to change when all the renewal data of un-to said secondary controller are updated behind, and said forward control unit becomes use impossible according to disaster etc., according to said duplex writing condition stored in said secondary controller The computing system characterized by making existence of data missing detectable.
[Claim 5] In a computing system according to claim 4 said forward control unit and said secondary controller In case said duplex writing inequality condition is made to change, the time of day made into said duplex writing inequality condition If said substorage is in the duplex writing inequality condition when it records on said forward control unit and said substorage, respectively and said forward control unit becomes use impossible according to disaster etc. The computing system characterized by performing data recovery from the backup file corresponding to the time of day which changed into said duplex writing inequality condition.
[Claim 6] The time of day transmitted to said secondary controller in a computing system according to claim 2 or 5 along with said positioning information or said duplex writing inequality condition is a computing system characterized by being given from said central processing unit as light processing demand time of day.
[Claim 7] The time of day transmitted to said secondary controller in a computing system according to claim 2 or 5 along with said positioning information or said duplex writing inequality condition is a computing system characterized by using the internal time amount of said forward control unit as time of day which had the light demand from said central processing unit.
[Claim 8] It is the computing system characterized by transmitting to said secondary controller only when it becomes clear in a computing system according to claim 2 that said positional information is light system access.
[Claim 9] It is the computing system characterized by what is notified to said secondary controller only when it becomes clear in a computing system according to claim 5 that said duplex writing inequality condition is light system access.
[Claim 10] It is the computing system characterized by said secondary controller reporting error termination to said central processing unit when there is access from said central processing unit in a storage system according to claim 4 to the substorage which is in said duplex writing inequality condition.

[Detailed Description of the Invention]
[0001]
[Field of the Invention] This invention relates to the computing system which consists of a forward storage system which connects with a central processing unit and consists of storage and a control unit, and a remote substorage system linked to the forward storage system. It is related with the computing system which improved the data guarantee approach at the time of performing remote duplex writing between the storage under said forward storage system, i.e., forward storage, and the storage under said substorage system, i.e., substorage, in more detail.
[0002]
[Description of the Prior Art] Through the central processing unit, data transfer is performed between control devices and the technique, i.e., remote duplex writing, of realizing duplex writing between different control devices is indicated by USP5155845. With this conventional technique, if a forward control device receives the command which specifies forward/** from a

central processing unit, the data memorized to the forward store will be copied to a substore. While storing light data in a forward store to the light demand to a forward control device from a central processing unit, the light data concerned are transmitted to a secondary controller, and light completion is reported to it and coincidence to a central processing unit after that. This is called a synchronous copy. Remote duplex writing is realized by the above actuation. When the data of a forward store become access impossible according to disaster, a failure, etc. by performing remote duplex writing, it becomes possible to succeed business with a substore.
[0003]
[Problem(s) to be Solved by the Invention] By the above-mentioned synchronous copy, since the time amount concerning the data transfer from a forward control unit to a secondary controller becomes long when the connection distance between a forward control unit and a secondary controller is as long as hundreds of km, the access engine performance to a central processing unit deteriorates. For this solution, before transmitting light data to a secondary controller, light completion is reported to the central processing unit, and the asynchronous copy method which transmits light data to a secondary controller behind can be considered. However, in an asynchronous copy method, the condition of being light unsettled produces data in a substore in light ending by a certain moment at a forward store. When a forward control device serves as use impossible according to a failure, disaster, etc. at this time, there is no means to recognize whether data have disappeared or not in a substore. For this reason, there is a trouble that it becomes impossible to judge whether substorage is usable. Then, the first purpose of this invention is to offer the computing system which can recognize with substorage whether data have disappeared or not, when an asynchronous copy method is applied and a forward control device becomes use impossible according to a failure, disaster, etc.
[0004] When it has been recognized that data have disappeared, it is necessary to recover the data which disappeared. If the data in a certain time are periodically backed up on the tape and there is need as an approach of recovering the data which disappeared, data will be recovered from backup data, and the time has a method of returning advance of a job. In order to perform data recovery by this approach, the time of day which disappeared must understand data. Then, the second purpose of this invention has data which disappeared in offering the computing system which can recognize the time of day by which the light was carried out from the central processing unit to forward storage with a secondary controller.
[0005]
[Means for Solving the Problem] In the computing system which consists of a forward storage system which connects this invention to a central processing unit, and consists of storage and a control unit in the 1st viewpoint, and a remote substorage system linked to the forward storage system the control unit, i.e., the forward control unit, under said forward storage system When a light demand is received from said central processing unit, after transmitting the positional information to said substore, without transmitting light data, light completion is reported to said central processing unit. Light processing to said substorage is performed at a suitable next stage., the control unit, i.e., the secondary controller, under said substorage system When said positional information transmitted from said forward control unit is held, light processing of being behind actual is performed, said positional information is canceled and said forward control unit becomes use impossible according to disaster etc. The computing system characterized by making existence of the data which disappeared detectable by said positional information stored in said secondary controller is offered. In the computing system by the 1st viewpoint of the above, since it is an asynchronous copy method, the access engine performance to a central

processing unit can be improved. Moreover, since the positional information transmitted from the forward control unit is stored with the secondary controller, when a forward control unit becomes use impossible according to a failure, disaster, etc., it can recognize with substorage whether data have disappeared or not by said positional information.

[0006] In the 2nd viewpoint, this invention is set to the computing system of the above-mentioned configuration. Said forward control unit To said positional information, in addition, by transmitting the time of day which had the light demand from said central processing unit to said secondary controller When said secondary controller is made to memorize and said forward control unit becomes use impossible according to disaster etc., the renewal sequence of data of said central processing unit The backup file which should be recovered is determined from the time of day stored in said secondary controller, and the computing system characterized by performing data recovery is offered. In the computing system by the 2nd viewpoint of the above, when said forward control unit becomes use impossible according to disaster etc., the time of day stored in the secondary controller shows at which time the contents of forward storage and substorage became an inequality. Therefore, in case business is taken over by the substore system, the data of proper back up time can be recovered.

[0007] In the 3rd viewpoint, when this invention has access in the field corresponding to said positional information from said central processing unit to said secondary controller in the computing system of the above-mentioned configuration, said secondary controller offers the computing system characterized by reporting error termination to said central processing unit. In the computer system by the 3rd viewpoint of the above, it can prevent accessing an old generation's data.

[0008] In the forward storage system which connects this invention to a central processing unit, and consists of storage and a control unit in the 4th viewpoint, and the computing system which consists of remote substorage systems linked to said forward storage system the control unit under said forward storage system, i.e., a forward control unit, and the control unit, i.e., the secondary controller, under said substorage system Each has managed two conditions, i.e., a duplex writing coincidence condition, and a duplex writing inequality condition as a condition [writing / duplex]. Said forward control unit If the forward storage for a light is in a duplex writing coincidence condition when a light demand is received from said central processing unit Make said each duplex writing condition of said forward storage and said substorage change in the duplex writing inequality condition, and light data report light completion to said substorage to said central processing unit, without transmitting. Light processing to said substorage is performed at a suitable next stage. Further When a duplex writing coincidence condition is made to change when all the renewal data of un-to said secondary controller are updated behind, and said forward control unit becomes use impossible according to disaster etc., according to said duplex writing condition stored in said secondary controller The computing system characterized by making existence of data missing detectable is offered. In the computing system by the 4th viewpoint of the above, since it is an asynchronous copy method, the access engine performance to a central processing unit can be improved. Moreover, as a duplex writing condition, since the condition of coincidence or an inequality is managed with a forward control unit and a secondary controller for every storage, when a forward control unit becomes use impossible according to a failure, disaster, etc., it can recognize with substorage whether data have disappeared or not according to said duplex writing condition.

[0009] In the 5th viewpoint, this invention is set to the computing system of the above-mentioned configuration. Said forward control unit and said secondary controller In case said

duplex writing inequality condition is made to change, the time of day made into said duplex writing inequality condition If said substorage is in the duplex writing inequality condition when it records on said forward control unit and said substorage, respectively and said forward control unit becomes use impossible according to disaster etc. The computing system characterized by performing data recovery from the backup file corresponding to the time of day which changed into said duplex writing inequality condition is offered. In the computing system by the 5th viewpoint of the above, when said forward control unit becomes use impossible according to disaster etc., the time of day stored in the secondary controller shows at which time the contents of forward storage and substorage became an inequality. Therefore, in case business is taken over by the substore system, the data of proper back up time can be recovered.

[0010] In the 6th viewpoint, the time of day when this invention is transmitted to said secondary controller in the computing system of the above-mentioned configuration along with said positioning information or said duplex writing inequality condition offers the computing system characterized by being given from said central processing unit as light processing demand time of day. An error with external time amount can be abolished in the computing system by the 6th viewpoint of the above.

[0011] In the 7th viewpoint, the time of day when this invention is transmitted to said secondary controller in the computing system of the above-mentioned configuration along with said positioning information or said duplex writing inequality condition offers the computing system characterized by using the internal time amount of said forward control unit as time of day which had the light demand from said central processing unit. In the computing system by the 7th viewpoint of the above, since it is not necessary to give time of day from a central processing unit, a configuration can be simplified.

[0012] In the 8th viewpoint, this invention offers the computing system characterized by said positional information transmitting to said secondary controller only when it becomes clear that it is light system access in the computing system of the above-mentioned configuration. In the computing system by the 8th viewpoint of the above, although positional information is transmitted only at the time of light system access, since it is only at the time of light system access, that the inequality of data happens with a forward store and a substore can avoid a useless communication link.

[0013] In the 9th viewpoint, in the computing system of the above-mentioned configuration, this invention offers the computing system characterized by what is notified to said secondary controller, only when it becomes clear that said duplex writing inequality condition is light system access. In the computing system by the 8th viewpoint of the above, although a duplex writing inequality condition is notified only at the time of light system access, since it is only at the time of light system access, that the inequality of data happens with a forward store and a substore can avoid a useless communication link.

[0014] In the 10th viewpoint, when this invention has access from said central processing unit in the storage system of the above-mentioned configuration to the substorage which is in said duplex writing inequality condition, said secondary controller offers the computing system characterized by reporting error termination to said central processing unit. In the computer system by the 10th viewpoint of the above, it can prevent accessing an old generation's data.

[0015]

[Embodiment of the Invention]

- The configuration of the computing system applied to the 1st operation gestalt of this invention at the 1st operation gestalt- drawing 1 is shown. In addition, in the following explanation,

although forward/** is distinguished by a of a sign, and b, when not distinguishing forward/**, a of a sign and b may be excluded.

[0016] This computing system 1 consists of subcomputing system 140b which is connected to forward computing system 140a and its forward computing system 140a, and is in remoteness. Said forward computing system 140a consists of forward central processing unit 100a and forward storage system 170a. Said subcomputing system 140b consists of subcentral processing unit 100b and forward storage system 170b. Said forward storage system 170a consists of forward control unit 110a and forward storage 130a. Said substorage system 170b consists of secondary controller 110b and substorage 130b.

[0017] Although determined by forward / sub**, and the user, generally the store employed at the time of usual is made forward, and the store which backs it up is sub**(ed). And suppose that forward [of a control unit 110, a central processing unit 100, and a computing system 140] and ** are decided on expedient or on logic by forward [of storage 130], or **. Although forward storage and substorage may be intermingled in one control unit 110 subordinate, in this case, to forward storage, that control unit 110 is a forward control unit, and turns into a secondary controller to substorage.

[0018] The control unit 110 is connected with a central processing unit 100 and other control units 110 by the link 150. The link control mechanism 120 performs the change of the link 150 which should be used etc. A cache 170 is memory which stores temporarily the data led from the data or the store 130 by which the light was carried out from the central processing unit 100. If the data which had access from the central processing unit 100 exist on a cache 170, access to storage 130 will not be performed. A processor 160 controls the data transfer between a store 130, a cache 170, a central processing unit 100, and other control devices 110. A directory 180 is the memory for storing the management information of a cache 170. A processor 160 performs a hit mistake judging etc. by accessing a directory 180.

[0019] A hit mistake judging means the processing which judges (a mistake) for whether it is that the data made into the purpose exist on a cache 170 from the location of the data which had the access request from the central processing unit 100, i.e., the store address, a data address, and a data length (hit).

[0020] A store 130 is a magnetic disk drive. However, a magnetic tape unit, an optical disk unit, etc. may be used. Drawing 2 is the explanatory view of a magnetic disk drive. This magnetic disk drive 250 consists of the heads 230 and the control unit interfaces 220 for carrying out read/write of two or more disks 240 which record data, and the data on a disk 240. While a disk 240 rotates one time, the record unit of the shape of a circle with an accessible head 230 is called a truck 200. Two or more trucks 200 exist on a disk 240. Each truck is divided into the field of the existing regular magnitude which is called a sector 260, and the sector serves as a smallest unit (slot) of a data access. When a certain truck 200 becomes a candidate for access, a head 230 is moved to the location which can carry out read/write of the truck 200. This actuation is called seeking. In this disk unit 250, two or more heads 230 move to coincidence. A number is assigned to ascending order from the top, and a head 230 calls it a head number. While a disk 240 rotates one time, the assembly of the truck 200 where all the heads 230 are accessible is called a cylinder 210. A number is assigned to the cylinder 210 by ascending order from the inside of a disk 240, and it is called a cylinder number.

[0021] The hit mistake judging table is stored in the directory 180. Drawing 3 is the example of structure of a hit mistake judging table. This hit mistake judging table 340 consists of next tables. Device-address conversion table 300: Having an entry every disk unit 250, each entry stores the

pointer to the cylinder number conversion table 310 corresponding to the disk unit concerned. Having an entry for every cylinder of the disk unit of 310:1 cylinder number conversion table, each entry stores the pointer to the track number conversion table 320 corresponding to the cylinder concerned.
Having an entry for every truck of the cylinder of 320:1 track number conversion table, each entry stores the pointer to the slot management table 330 corresponding to the truck concerned.
It is the cache management information of the truck of 330:1 slot management table, and has the following information.
Data existence map 331: It is the information which shows which sector on a truck exists on a cache 170.
Cache address 332: When the sector on a truck exists on a cache 170, it is the address information in which the data is stored.
Pointer 333: It is the information used when performing queue management etc.
Time of day 334: It is the information for managing the time of day which had the light in forward storage 130a to the slot concerned by the secondary controller side 110b side.
RVOL dirty condition 335: The slot concerned is the information which shows whether it is renewal of un-to substorage 130b.
Track address 336: It is the address of the truck concerned.
Head address 337: It is the address of the head which accesses the truck concerned.
[0022] The device address conversion table 300, the cylinder number conversion table 310, the track number conversion table 320, and the slot management table 330 are followed, based on the positioning information, i.e., the storage address for access, specified from the central processing unit 100, a cylinder number, and a track number, if the bit to which the data existence map 331 corresponds is ON, it will judge with a hit, and if off, it will judge with a mistake. Furthermore, in performing data transfer, it computes the cache address used as the candidate for a transfer by adding the offset corresponding to the target sector to the cache address 332.
[0023] In order to realize remote duplex writing by the asynchronous copy, it has the RVOL dirty management table which manages the data (this is called RVOL dirty data) of the renewal of un-to substore 130b in a directory 180.
[0024] Drawing 4 is the block diagram of a RVOL dirty management table. This RVOL dirty management table 400 has the entry of every storage VOL1 and VOL2 and --. The RVOL information 401 and the RVOL dirty list head 402 are stored in one entry. To the RVOL information 401, the secondary controller 110b address when making the storage concerned forward and the substorage 130b address are held. A special value is stored when the storage 130 concerned has not constructed the pair of forward/**. The RVOL dirty list head 402 points out the RVOL dirty list 410 of [for managing the slot of renewal of un-]. A NULL value is stored when the data of renewal of un-do not exist. The RVOL dirty list 410 makes the list structure the slot management table 330 of the renewal slot of un-using the pointer 333 in the slot management table 330.
[0025] Registration of the RVOL dirty management table 400 is performed when proved that they are the time of light system access to forward storage 130a from a central processing unit 100, or light system access. A central processing unit 100 is accessed to forward storage 130a according to the defined protocol. Although various things about the protocol exist, generally, for example in the mainframe system centering on a mainframe, the CKD protocol is used. A positioning command including the information which expresses the access gestalt (a lead or light system access) other than positioning information with this CKD protocol in advance of a

data-access command is published. Therefore, when whether it is light system access receives a positioning command, in order to become clear, the RVOL dirty management table 400 is registered at this time. Moreover, in the workstation or the personal computer, the FBA protocol which includes positioning information in a data-access command is used. In the case of this FBA protocol, registration of the RVOL dirty management table 400 is performed after data-access command receipt.

[0026] Drawing 5 is the flow chart of registration processing of the RVOL dirty management table 400. At step 510, it judges whether it is light system access as mentioned above, if it is not light system access, processing will be ended, and if it becomes clear that it is light system access, it will progress to step 520. At step 520, it is confirmed by asking for the entry of the RVOL dirty management table 400 from the address of the store 130 for access, and referring to the RVOL information 401 on the entry concerned whether the store 130 for access forms the pair of forward/**. If the pair is not formed, processing is ended and the pair is formed, it will progress to step 530. At step 530, it asks for the address of secondary controller 110b from the RVOL information 401, and positioning information is transmitted to the secondary controller 110b. Thereby, it can notify that non-reflected data exist in substore 130b to secondary controller 110b. The slot for a light is registered into the RVOL dirty management table 400 at step 540. That is, the new slot management table 330 is inserted in front of the slot management table 330 which the RVOL dirty list header 402 points out.

[0027] In addition, in said step 530, the processing which transmits positioning information to secondary controller 110b is good to carry out in parallel to the processing which receives light data from a central processing unit, in order to reduce the overheads on appearance. For example, if a CKD protocol explains, after receiving positioning information with a positioning command, a child job will be generated, positioning information transfer processing will be performed, and processing of the light command chained with a positioning command will be performed by the self-job. By the self-job, after the command chain concerned is completed, and waiting for completion of a child job, the completion of light processing is reported to a central processing unit 100.

[0028] Moreover, in said step 530, the current time of day other than positioning information may be transmitted to secondary controller 110b. Thereby, it can investigate now behind at which time data disappeared by the secondary controller 110b side. Furthermore, in case the present time of day is transmitted, the internal time amount of forward control unit 110a may be transmitted, but in order to abolish an error with external time amount, in the positioning command parameter from a central processing unit 100, time of day is stored and the time of day may be transmitted. For example, using the RVOL dirty information 600 shown in drawing 6 R> 6, positioning information and time of day are transmitted to secondary controller 110b from forward control unit 110a. This RVOL dirty information 600 consists of the command code 601 showing being a transfer of RVOL dirty information, the substore address 610, dirty data positional information 620, and time of day 630.

[0029] Drawing 7 is the flow chart of the deletion of the RVOL dirty management table 400. In addition, in order to realize remote duplex writing of an efficient asynchronous copy method, the asynchronous duplex writing job which performs a RVOL light to under exclusive contract shall exist in the mounted forward store 130a correspondence. For every fixed period, this asynchronous duplex writing job is a job which repeats a run and sleep, and performs deletion of the RVOL dirty management table 400 during a run. At step 710, since RVOL dirty data do not exist with reference to the RVOL dirty list header 402 of the RVOL dirty management table 400

of forward store 130a corresponding to the asynchronous duplex writing job under activation if the value becomes NULL, processing is ended. If the value of the RVOL dirty list header 402 is not NULL, since RVOL dirty data exist, it progresses to step 720. At step 720, the light command to substorage 130b is generated from the RVOL information 401 on the RVOL dirty management table 400, and the information on the slot management table 330 concerned, and RVOL light processing is performed. In addition, the generation method of a light command is indicated by USP555845. At step 730, the slot management table 330 which carried out light completion is deleted from the RVOL dirty list 410. If processing is completed, it sleeps for a while.

[0030] The above is processing by the side of forward control unit 110a. Next, the processing by the side of secondary controller 110b is explained.

[0031] The main processings of secondary controller 110b are holding the RVOL dirty information 600 after the RVOL dirty information 600 is transmitted from forward control-device 110a until the light of the data concerned is carried out. In order to realize this processing, it has the same table structure fundamentally also by secondary controller 110b with the RVOL dirty data control table 400 shown in drawing 4 . The slot management table 330 connected to the RVOL dirty list 410 expresses a non-reflected slot to substorage 130b. In addition, the RVOL information 401 is not used in secondary controller 110b.

[0032] Registration to the RVOL dirty data control table 400 is performed when the RVOL dirty information 600 has been transmitted from forward control-device 110a. That is, it judges whether it is a transfer of the RVOL dirty data information 600 by command code 601, and if it is a transfer of the RVOL dirty information 600, it will ask for the corresponding RVOL dirty management table 400 of substore 130b and the corresponding RVOL dirty list 410 from the store address 610. Moreover, the activation sushi and target slot management table 330 is obtained from the dirty data positional information 620. [judging / hit mistake] And while registering the slot management table 330 into said RVOL dirty list 410, let the RVOL dirty condition 335 of the slot management table 330 be "those with dirty."

[0033] Deletion of the slot management table 330 is performed when the light to the object slot from forward control unit 110a is performed. That is, a hit mistake judging is performed based on positioning information to the light demand from forward control unit 110a, if it is "with dirty", after performing light processing with reference to the RVOL dirty condition 335 of the obtained slot management table 330, the slot management table 330 concerned is deleted from the RVOL dirty list 410, and the RVOL dirty information 335 is changed into "-dirty-less ".

[0034] The time of day when the positional information of the renewal data of un-to substore 130b and its data became renewal of un-by the above will be held, respectively by forward control unit 110a and secondary controller 110b.

[0035] The disappearance data list for recording the positional information of disappearance data and disappearance time of day on drawing 8 is shown. This disappearance data list 1100 consists of the store address 1110, 1120 disappearance data, and the data address 1130 for every disappearance data and the time of day 1140 which disappeared.

[0036] In addition, if said number of disappearance data exceeds a predetermined threshold, the storage 130 whole becomes an invalid and you may make it recover the storage 130 whole. If it carries out like this, it can prevent the disappearance data list 1100 becoming huge. As for said threshold, it is desirable that it can set up from a customer engineer etc.

[0037] Secondary controller 110b creates the disappearance data list 1100 with the following procedure to the read-out demand of the disappearance data list 1100 from a central processing

unit 100.
(1) Search the corresponding RVOL dirty management table 400 from the store address with a demand.
(2) If the value of the RVOL dirty list head 402 of the searched RVOL dirty management table 400 is NULL, since disappearance data do not exist, record "0" on 1120 disappearance data.
(3) If the value of the RVOL dirty list head 402 has pointed out the RVOL dirty list 410, store the track address 336 of the slot management table 330 linked to the RVOL dirty list 410, and the head address 337 in a data address 1130, and store time of day 334 in time of day 1140. The number of disappearance data is counted repeatedly, following the RVOL dirty list 410 for this.
(4) Store the counted number of disappearance data in 1120 disappearance data.
(5) Store the specified storage address in the storage address 1110.
[0038] If the disappearance data list 1100 is completed by the above, secondary controller 110b will return the disappearance data list 1100 to a central processing unit 100. Then, a customer engineer can investigate the existence of disappearance data by reading the disappearance data list 1100. And when data have disappeared, data can be recovered from a data address 1130. Moreover, the oldest thing of time of day 1140 is searched, and it also becomes possible from the generation's backup file to recover data.
[0039] Drawing 9 is a flow chart showing the access propriety judging processing 800 in secondary controller 110b when access of a forward control unit has become impossible. At step 810, the hit mistake judging of the slot for access is performed, and the slot management table 330 is obtained. At step 820, the data of the slot concerned judge whether it is un-reflected from the RVOL dirty condition 335 of said slot management table 330 to substorage 130b. If not reflected, since data will have disappeared, it progresses to step 830. If reflected, since data will not have disappeared, it progresses to step 840.
[0040] Step 830 reports abnormal termination to a central processing unit 100. Access is permitted at step 840.
[0041] Although explanation of the 1st operation gestalt is finished above, it is as follows when the main point of the 1st operation gestalt is summarized here.
(1) Transmit positioning information and time of day to secondary controller 110b before actually performing light processing from forward control unit 110a. Secondary controller 110b holds said positioning information and said time of day until a light is actually performed.
(2) When forward store 130a becomes access impossible according to a failure, disaster, etc., judge the existence of disappearance data from secondary controller 110b. When there are disappearance data, the backup file which should be recovered is determined from the time of day of the disappearance data.
(3) When there is access to disappearance data, don't permit access but report abnormal termination.
[0042] Writing [remote duplex] of the conventional asynchronous copy method, when data are not recovered although disappearance data exist since existence of the data which disappeared cannot be judged, data transformation occurs. Since recovery becomes useless and data return several generations ago on the other hand when data are recovered, although disappearance data do not exist, cost starts extremely. Moreover, the storage 130 whole must be recovered and effectiveness is bad. On the other hand, writing [remote duplex] of the asynchronous copy method by the operation gestalt of the above 1st, before starting operation to the site of sub**, if inharmonious, after it checks, and carrying out data recovery of whether the contents of forward store 130a and substore 130b are in agreement first from the backup file of suitable time of day,

operation is started. Therefore, said data transformation and useless data recovery are avoidable. By the above, even when access to forward storage 110a becomes impossible according to disaster, a failure, etc., business can be succeeded to the site of sub** using substorage 130b, and, moreover, system stop time amount is made to min.

[0043] - 2nd operation gestalt - With said 1st operation gestalt, since positioning information was transmitted to secondary controller 110b for every (every CC chain) one light demand from a central processing unit 100, the count of a communication link between forward control unit 110a and secondary controller 110b increases, and the engine performance falls. With the 2nd operation gestalt, in order to solve this, only when the contents of forward storage 130a and substorage 130b are coincidence or an inequality, that is notified to secondary controller 110b, and the count of a communication link between forward control unit 110a and secondary controller 110b is reduced.

[0044] Storage 130 takes two conditions of duplex writing coincidence or a duplex writing inequality. Here, it is shown that duplex writing coincidence is in the condition whose contents of forward storage 130a and substorage 130b corresponded. Moreover, a duplex writing inequality shows that the contents of forward storage 130a and substorage 130b are in the condition of an inequality. These duplex writing conditions are stored in the RVOL information 401 on the RVOL dirty management table 400 of forward control-device 130a and secondary controller 130b. Moreover, the time of day used as a duplex writing inequality is also stored in the RVOL information 401.

[0045] Drawing 10 is the flow chart of the inequality-ized status-change processing 900 which carries out a state transition to a duplex writing inequality from duplex writing coincidence. At step 910, it confirms whether to be light system access. If it is except light system access, since modification of a duplex writing condition will not be generated, processing is ended. If it is light system access, it will progress to step 920. At step 920, it judges whether the storage 130 for access forms the pair of forward/** from the RVOL information 401. If the pair is not formed, since modification of a duplex writing condition is not generated, processing is ended. When the pair is formed, it progresses to step 930. At step 930, a current duplex writing condition judges whether it is duplex writing coincidence from the RVOL information 401. If it is a duplex writing inequality, since modification of a duplex writing condition will not be generated, processing is ended. If it is duplex writing coincidence, it will progress to step 940. At step 940, it notifies having become a duplex writing inequality to secondary controller 110b. It is desirable to perform processing of this notice and light processing from a central processing unit 100 in parallel from a viewpoint of effectiveness. At step 950, the status change of the RVOL information 401 is carried out to a duplex writing inequality.

[0046] In addition, not only the notice of a duplex writing inequality but its time of day may be notified to secondary controller 110b by the above-mentioned inequality-ized status-change processing 900. In this case, the time of day which received light system access from the central processing unit 100 is transmitted to secondary controller 110b, and the time of day concerned is recorded also on the RVOL information 401. In addition, the internal time of day of forward control unit 110a is sufficient as this time of day, and the time of day given from a central processing unit 100 is sufficient as it.

[0047] Drawing 11 is the flow chart of the identification status-change processing 1000 which carries out a state transition to duplex writing coincidence from a duplex writing inequality. This identification status-change processing 1000 is performed when the asynchronous duplex writing job explained with the 1st operation gestalt carries out the light of the renewal data of un-to

substore 130b. At step 1000, as the 1st operation gestalt explained, the light of the renewal data of un-is carried out to substorage 130b. At step 1010, the slot management table 330 which carried out light completion as the 1st operation gestalt explained is removed from the RVOL dirty list 410. At step 1020, it judges whether the renewal data of un-to substore 130b exist by following the RVOL dirty list 410. If the slot management table 330 is still connected to the RVOL dirty list 410, since it is still a duplex writing inequality, processing is ended. If the slot management table 330 was not connected to the RVOL dirty list 410, since the duplex writing inequality was canceled, it progresses to step 1030. At step 1030, it notifies having become duplex writing coincidence to secondary controller 110b. At step 1040, the status change of the RVOL information 401 is carried out to duplex writing coincidence.

[0048] Drawing 12 is the instantiation Fig. of the notice information of a status change used by the notice of the status change to secondary controller 110b from forward control unit 110a. The command code 1301 which shows that this notice information 1300 of a status change is a notice command of a status change, the substorage address 1310 which is applicable, the time of day 1320 which performed the status change, and the condition code 1330 showing the code of a new condition are stored.

[0049] The above is processing by the side of forward control unit 110a. Next, the processing by the side of secondary controller 110b is explained.

[0050] In secondary controller 110b, receipt of the notice information 1300 of a status change changes the RVOL information 401 corresponding to substorage 130b which is applicable into the condition of having been specified by the condition code 1330. In addition, the RVOL dirty list 410 is not used in secondary controller 110b.

[0051] The status information 1200 of principal and vice for notifying the duplex writing condition of storage 130 to drawing 13 at a central processing unit 100 is shown. This status information 1200 of principal and vice consists of the storage address 1210, a duplex writing condition 1220 of expressing duplex writing coincidence or a duplex writing inequality, and status-change time of day 1230 showing the time of day used as a duplex writing inequality.

[0052] Secondary controller 110b creates the status information 1200 of principal and vice with the following procedure to a read-out demand of the status information of principal and vice from a central processing unit 100.

(1) From the specified store address, search the corresponding RVOL dirty management table 400, from the RVOL information 401, acquire a duplex writing condition and store in the duplex writing condition 1220.

(2) If it is a duplex writing inequality, store in the status-change time of day 1230 the time of day recorded on the RVOL information 401.

If the status information 1200 of principal and vice is completed by the above, secondary controller 110b will return the status information 1200 of principal and vice to a central processing unit 100. Then, a customer engineer can check a duplex writing condition by reading the status information 1200 of principal and vice. And if it is a duplex writing inequality, a backup file with the nearest generation can be selected out of the time of day used as an inequality, and data recovery can be performed. The approach of this 2nd operation gestalt cannot perform each data recovery, or when unnecessary, it is useful.

[0053] If there is an access request from a central processing unit 100 to substorage 130b of a duplex writing inequality, the secondary controller 110b asks for the RVOL dirty management table 400 which corresponds from the storage address with a demand, and when the pair of whether the storage 130 concerned has constructed the pair of forward/**, and forward/** is

constructed from the RVOL information 401, it will investigate whether it is a duplex writing inequality. And in the case of a duplex writing inequality, access is refused in order to prevent transmitting an old generation's data.
[0054] Although explanation of the 2nd operation gestalt is finished above, it is as follows when the main point of the 2nd operation gestalt is summarized here.
(1) Hold the duplex writing condition of expressing coincidence/inequality of the contents of forward and the substorage 130, to a control unit 110.
(2) A duplex writing condition serves as a duplex writing inequality, when the renewal data of un-are made into substore 130b, and when the renewal data of un-are lost, it serves as duplex writing coincidence.
(3) When access to forward storage 130a becomes impossible according to a failure or disaster, a duplex writing condition can be read from secondary controller 110b, and data can be recovered from the nearest generation's backup file from the time of day used as a duplex writing inequality.
(4) Access from the central processing unit 100 to substorage 130b used as a duplex writing inequality is refused.
[0055] or [that the contents of a customer engineer of forward store 130a and substore 130b correspond with the operation gestalt of the above 2nd before starting operation to the site of sub**] -- it confirms whether to be an inequality. And if inharmonious, operation will be started after recovering data from the backup file of a suitable stage. By the above, even when access to forward storage 110a becomes impossible according to disaster, a failure, etc., business can be succeeded to the site of sub** using substorage 130b, and, moreover, system stop time amount is made to min.
[0056] - difference - of the 1st operation gestalt and the 2nd operation gestalt -- (1) -- with the 1st operation gestalt, the response time of the light processing from a central processing unit 100 may deteriorate. On the other hand, with the 2nd operation gestalt, there is almost no response performance degradation.
(2) It is necessary to recover the whole storage with the 2nd operation gestalt to a thing recoverable to the data unit which disappeared with the 1st operation gestalt in recovery of the disappearance data after a failure and disaster.
[0057]
[Effect of the Invention] If the renewal data of un-to a substore are made, since it will notify to a secondary controller that the positioning information or an original and copy duplex writing condition is inharmonious from a forward control unit according to the computer system of this invention, when a forward store becomes use impossible according to a failure, disaster, etc., the existence of disappearance data can be judged from a secondary controller, and it becomes possible to cope with it appropriately.

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－325917

(43)公開日　平成９年(1997)12月16日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	12/16	３１０	7623－5Ｂ	Ｇ０６Ｆ	12/16	３１０Ｊ
	3/06	３０４			3/06	３０４Ｆ
		３０６				３０６Ｂ

審査請求　未請求　請求項の数10　ＯＬ　（全 14 頁）

(21)出願番号　特願平8－145562

(22)出願日　平成８年(1996)６月７日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　山神　遠司
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　中村　勝憲
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内

(72)発明者　山本　彰
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(74)代理人　弁理士　有近　紳志郎

(54)【発明の名称】　計算機システム

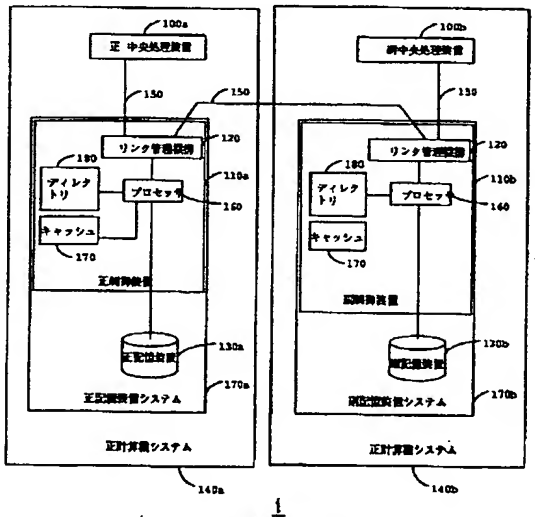
(57)【要約】

【課題】　非同期コピー方式を適用した場合において、障害や災害等により正制御装置が使用不能となった時に、データが消失しているかどうかを副記憶装置で認識できる計算機システムを提供する。

【解決手段】　中央処理装置１００からのライト要求に対して、正制御装置１１０ａはライト対象データの位置情報のみを副制御装置１１０ｂへ転送する。副制御装置１１０ｂでは、実際にデータがライトされるまで、前記位置情報を保持しておく。障害や災害により正記憶装置１３０ａがアクセス不能になった場合、副制御装置１１０ｂから位置情報を読み出すことにより、消失データの有無およびその位置や時刻を判別する。

【効果】　障害や災害等により正記憶装置が使用不能となった時に、消失データの有無を副制御装置から判断でき、適切に対処することが可能になる。

図1



1
計算機システム

1

【特許請求の範囲】

【請求項１】　中央処理装置に接続しかつ記憶装置と制御装置から構成される正記憶装置システムと、その正記憶装置システムに接続する遠隔の副記憶装置システムとから構成される計算機システムにおいて、

前記正記憶装置システム下の制御装置すなわち正制御装置は、前記中央処理装置からライト要求を受けた際、前記副記憶装置へはライトデータは転送せずに、その位置情報を転送した後、前記中央処理装置に対してはライト完了を報告して、後の適当な時期に前記副記憶装置に対するライト処理を実行し、

前記副記憶装置システム下の制御装置すなわち副制御装置は、前記正制御装置から転送された前記位置情報を保持しておき、後に実際のライト処理が実行された時に、前記位置情報を破棄し、

災害等により前記正制御装置が使用不能となったときに、前記副制御装置に格納されている前記位置情報によって、消失したデータの有無を検出可能としたことを特徴とする計算機システム。

【請求項２】　請求項１に記載の計算機システムにおいて、前記正制御装置は、前記位置情報に加えて、前記中央処理装置からライト要求のあった時刻を前記副制御装置に転送することによって、前記中央処理装置のデータ更新順序を、前記副制御装置に記憶させ、

災害等により前記正制御装置が使用不能になったときに、前記副制御装置に格納されている時刻から、回復すべきバックアップファイルを決定し、データ回復を行うことを特徴とする計算機システム。

【請求項３】　請求項１に記載の計算機システムにおいて、前記中央処理装置から前記副制御装置に対して、前記位置情報に対応する領域にアクセスがあった場合に、前記副制御装置は、エラー終了を前記中央処理装置に報告することを特徴とする計算機システム。

【請求項４】　中央処理装置に接続しかつ記憶装置と制御装置から構成される正記憶装置システムと、前記正記憶装置システムに接続する遠隔の副記憶装置システムから構成される計算機システムにおいて、

前記正記憶装置システム下の制御装置すなわち正制御装置と、前記副記憶装置システム下の制御装置すなわち副制御装置は、それぞれが、二重書きの状態として２つの状態、すなわち二重書き一致状態および二重書き不一致状態を管理しており、

前記正制御装置は、前記中央処理装置からライト要求を受けた際、ライト対象の正記憶装置が二重書き一致状態であれば、前記正記憶装置および前記副記憶装置のそれぞれの前記二重書き状態を二重書き不一致状態に遷移させ、前記副記憶装置へはライトデータは転送せずに、前記中央処理装置に対してはライト完了を報告して、後の適当な時期に前記副記憶装置に対するライト処理を実行し、さらに、後に前記副制御装置に対する未更新データ

2

が全て更新された時に二重書き一致状態に遷移させ、

災害等により前記正制御装置が使用不能となった時に、前記副制御装置に格納されている前記二重書き状態によって、データ消失の有無を検出可能としたことを特徴とする計算機システム。

【請求項５】　請求項４に記載の計算機システムにおいて、前記正制御装置および前記副制御装置を、前記二重書き不一致状態に遷移させる際に、前記二重書き不一致状態とした時刻を、前記正制御装置および前記副記憶装置にそれぞれ記録し、

災害等により前記正制御装置が使用不能になった場合に、前記副記憶装置が二重書き不一致状態となっていれば、前記二重書き不一致状態となった時刻に対応するバックアップファイルからデータ回復を行なうことを特徴とする計算機システム。

【請求項６】　請求項２または請求項５に記載の計算機システムにおいて、前記位置付け情報あるいは前記二重書き不一致状態に付随して前記副制御装置に転送される時刻は、ライト処理要求時刻として、前記中央処理装置から与えられることを特徴とする計算機システム。

【請求項７】　請求項２または請求項５に記載の計算機システムにおいて、前記位置付け情報あるいは前記二重書き不一致状態に付随して前記副制御装置に転送される時刻は、前記中央処理装置からライト要求のあった時刻として、前記正制御装置の内部時間を使用することを特徴とする計算機システム。

【請求項８】　請求項２に記載の計算機システムにおいて、前記位置情報は、ライト系アクセスであることが判明したときのみ、前記副制御装置に転送することを特徴とする計算機システム。

【請求項９】　請求項５に記載の計算機システムにおいて、前記二重書き不一致状態は、ライト系アクセスであることが判明したときのみ、前記副制御装置に通知することを特徴とする計算機システム。

【請求項１０】　請求項４に記載の記憶装置システムにおいて、前記中央処理装置から、前記二重書き不一致状態である副記憶装置に対してアクセスがあった場合に、前記副制御装置は、エラー終了を前記中央処理装置に報告することを特徴とする計算機システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、中央処理装置に接続しかつ記憶装置と制御装置から構成される正記憶装置システムと、その正記憶装置システムに接続する遠隔の副記憶装置システムとから構成される計算機システムに関する。さらに詳しくは、前記正記憶装置システム下の記憶装置すなわち正記憶装置と、前記副記憶装置システム下の記憶装置すなわち副記憶装置の間で遠隔二重書きを行なう際のデータ保証方法を改良した計算機システムに関する。

【0002】
【従来の技術】中央処理装置を介することなしに制御装置間でデータ転送を行ない、異なる制御装置間で二重書きを実現する技術、すなわち遠隔二重書きは、USP5155845に記載されている。この従来技術では、中央処理装置から正／副を指定するコマンドを正制御装置が受けとると、正記憶装置に記憶していたデータを副記憶装置にコピーする。それと同時に、中央処理装置から正制御装置へのライト要求に対して、ライトデータを正記憶装置に格納すると共に、当該ライトデータを副制御装置に転送し、その後、中央処理装置に対してライト完了を報告する。これを同期コピーと呼ぶ。以上の動作により、遠隔二重書きが実現される。遠隔二重書きを行なうことによって、災害や障害等によって、正記憶装置のデータがアクセス不能になった場合に、副記憶装置によって業務を引き継ぐことが可能になる。
【0003】
【発明が解決しようとする課題】上記の同期コピーでは、正制御装置と副制御装置の間の接続距離が数百ｋｍと長い場合には、正制御装置から副制御装置へのデータ転送にかかる時間が長くなるために、中央処理装置に対するアクセス性能が劣化する。この解決のためには、副制御装置に対してライトデータを転送する前にライト完了を中央処理装置に報告しておき、後に副制御装置に対してライトデータを転送する、非同期コピー方式が考えられる。しかし、非同期コピー方式では、ある瞬間では、正記憶装置にはデータをライト済みで、副記憶装置にはライト未済である状態が生じる。この時に障害や災害等により正制御装置が使用不能となると、副記憶装置では、データが消失しているかどうかを認識する手段がない。このために、副記憶装置が使用可能かどうかを判断できなくなる問題点がある。そこで、本発明の第一の目的は、非同期コピー方式を適用した場合において、障害や災害等により正制御装置が使用不能となった時に、データが消失しているかどうかを副記憶装置で認識できる計算機システムを提供することにある。
【0004】データが消失していることを認識できた場合、その消失したデータを回復する必要がある。消失したデータを回復する方法としては、ある時点でのデータを定期的にテープにバックアップしておき、必要があればバックアップデータからデータを回復して、その時点までジョブの進行を戻す方法がある。この方法でデータ回復を行なうには、データを消失した時刻が判らなくてはならない。そこで、本発明の第二の目的は、消失したデータが中央処理装置から正記憶装置に対してライトされた時刻を副制御装置で認識できる計算機システムを提供することにある。
【0005】
【課題を解決するための手段】第１の観点では、本発明は、中央処理装置に接続しかつ記憶装置と制御装置から構成される正記憶装置システムと、その正記憶装置システムに接続する遠隔の副記憶装置システムとから構成される計算機システムにおいて、前記正記憶装置システム下の制御装置すなわち正制御装置は、前記中央処理装置からライト要求を受けた際、前記副記憶装置へはライトデータは転送せずに、その位置情報を転送した後、前記中央処理装置に対してはライト完了を報告して、後の適当な時期に前記副記憶装置に対するライト処理を実行し、前記副記憶装置システム下の制御装置すなわち副制御装置は、前記正制御装置から転送された前記位置情報を保持しておき、後に実際のライト処理が実行された時に、前記位置情報を破棄し、災害等により前記正制御装置が使用不能となったときに、前記副制御装置に格納されている前記位置情報によって、消失したデータの有無を検出可能としたことを特徴とする計算機システムを提供する。上記第１の観点による計算機システムでは、非同期コピー方式であるから、中央処理装置に対するアクセス性能を向上できる。また、正制御装置から転送された位置情報を副制御装置で格納しておくから、障害や災害等により正制御装置が使用不能となった時に、データが消失しているかどうかを、前記位置情報により副記憶装置で認識することが出来る。
【0006】第２の観点では、本発明は、上記構成の計算機システムにおいて、前記正制御装置は、前記位置情報に加えて、前記中央処理装置からライト要求のあった時刻を前記副制御装置に転送することによって、前記中央処理装置のデータ更新順序を、前記副制御装置に記憶させ、災害等により前記正制御装置が使用不能になったときに、前記副制御装置に格納されている時刻から、回復すべきバックアップファイルを決定し、データ回復を行うことを特徴とする計算機システムを提供する。上記第２の観点による計算機システムでは、災害等により前記正制御装置が使用不能になったときに、副制御装置に格納された時刻により、どの時点で正記憶装置と副記憶装置の内容が不一致になったかが判る。よって、副記憶装置システムで業務の引き継ぎを行なう際、適正なバックアップ時刻のデータから回復することが出来る。
【0007】第３の観点では、本発明は、上記構成の計算機システムにおいて、前記中央処理装置から前記副制御装置に対して、前記位置情報に対応する領域にアクセスがあった場合に、前記副制御装置は、エラー終了を前記中央処理装置に報告することを特徴とする計算機システムを提供する。上記第３の観点による計算機システムでは、旧世代のデータにアクセスするのを防止することが出来る。
【0008】第４の観点では、本発明は、中央処理装置に接続しかつ記憶装置と制御装置から構成される正記憶装置システムと、前記正記憶装置システムに接続する遠隔の副記憶装置システムから構成される計算機システムにおいて、前記正記憶装置システム下の制御装置すなわ

ち正制御装置と、前記副記憶装置システム下の制御装置すなわち副制御装置は、それぞれが二重書きの状態として2つの状態すなわち二重書き一致状態および二重書き不一致状態を管理しており、前記正制御装置は、前記中央処理装置からライト要求を受けた際、ライト対象の正記憶装置が二重書き一致状態であれば、前記正記憶装置および前記副記憶装置のそれぞれの前記二重書き状態を二重書き不一致状態に遷移させ、前記副記憶装置へはライトデータは転送せずに、前記中央処理装置に対してはライト完了を報告して、後の適当な時期に前記副記憶装置に対するライト処理を実行し、さらに、後に前記副制御装置に対する未更新データが全て更新された時に二重書き一致状態に遷移させ、災害等により前記正制御装置が使用不能となった時に、前記副制御装置に格納されている前記二重書き状態によって、データ消失の有無を検出可能としたことを特徴とする計算機システムを提供する。上記第4の観点による計算機システムでは、非同期コピー方式であるから、中央処理装置に対するアクセス性能を向上できる。また、二重書き状態として、一致または不一致の状態を各記憶装置毎に、正制御装置および副制御装置で管理するから、障害や災害等により正制御装置が使用不能となった時に、データが消失しているかどうかを、前記二重書き状態により副記憶装置で認識することが出来る。

【0009】第5の観点では、本発明は、上記構成の計算機システムにおいて、前記正制御装置および前記副制御装置を、前記二重書き不一致状態に遷移させる際に、前記二重書き不一致状態とした時刻を、前記正制御装置および前記副記憶装置にそれぞれ記録し、災害等により前記正制御装置が使用不能になった場合に、前記副記憶装置が二重書き不一致状態となっていれば、前記二重書き不一致状態となった時刻に対応するバックアップファイルからデータ回復を行なうことを特徴とする計算機システムを提供する。上記第5の観点による計算機システムでは、災害等により前記正制御装置が使用不能になったときに、副制御装置に格納された時刻により、どの時点で正記憶装置と副記憶装置の内容が不一致になったかが判る。よって、副記憶装置システムで業務の引き継ぎを行なう際、適正なバックアップ時刻のデータから回復することが出来る。

【0010】第6の観点では、本発明は、上記構成の計算機システムにおいて、前記位置付け情報あるいは前記二重書き不一致状態に付随して前記副制御装置に転送される時刻は、ライト処理要求時刻として、前記中央処理装置から与えられることを特徴とする計算機システムを提供する。上記第6の観点による計算機システムでは、外部時間との誤差をなくすことが出来る。

【0011】第7の観点では、本発明は、上記構成の計算機システムにおいて、前記位置付け情報あるいは前記二重書き不一致状態に付随して前記副制御装置に転送される時刻は、前記中央処理装置からライト要求のあった時刻として、前記正制御装置の内部時間を使用することを特徴とする計算機システムを提供する。上記第7の観点による計算機システムでは、中央処理装置から時刻を与える必要がないため、構成を簡単化できる。

【0012】第8の観点では、本発明は、上記構成の計算機システムにおいて、前記位置情報は、ライト系アクセスであることが判明したときのみ、前記副制御装置に転送することを特徴とする計算機システムを提供する。上記第8の観点による計算機システムでは、ライト系アクセスのときのみ位置情報の転送を行うが、正記憶装置と副記憶装置でデータの不一致が起こるのはライト系アクセスのときのみであるため、無駄な通信を回避できる。

【0013】第9の観点では、本発明は、上記構成の計算機システムにおいて、前記二重書き不一致状態は、ライト系アクセスであることが判明したときのみ、前記副制御装置に通知することを特徴とする計算機システムを提供する。上記第8の観点による計算機システムでは、ライト系アクセスのときのみ二重書き不一致状態の通知を行うが、正記憶装置と副記憶装置でデータの不一致が起こるのはライト系アクセスのときのみであるため、無駄な通信を回避できる。

【0014】第10の観点では、本発明は、上記構成の記憶装置システムにおいて、前記中央処理装置から、前記二重書き不一致状態である副記憶装置に対してアクセスがあった場合に、前記副制御装置は、エラー終了を前記中央処理装置に報告することを特徴とする計算機システムを提供する。上記第10の観点による計算機システムでは、旧世代のデータにアクセスするのを防止することが出来る。

【0015】

【発明の実施の形態】

－第1の実施形態－

図1に、本発明の第1の実施形態にかかる計算機システムの構成を示す。なお、以下の説明では、正／副を符号のa，bにより区別するが、正／副を区別しないときには符号のa，bを省く場合もある。

【0016】この計算機システム1は、正計算機システム140aと、その正計算機システム140aに接続され且つ遠隔にある副計算機システム140bとから構成されている。前記正計算機システム140aは、正中央処理装置100aと、正記憶システム170aとから構成されている。前記副計算機システム140bは、副中央処理装置100bと、正記憶システム170bとから構成されている。前記正記憶システム170aは、正制御装置110aと、正記憶装置130aとから構成されている。前記副記憶システム170bは、副制御装置110bと、副記憶装置130bとから構成されている。

【0017】正／副は、ユーザによって決定されるが、

一般的には、平常時に運用する記憶装置を正とし、それをバックアップする記憶装置を副とする。そして、記憶装置１３０の正または副によって、便宜上あるいは論理上、制御装置１１０、中央処理装置１００、計算機システム１４０の正および副を決めることとする。一つの制御装置１１０配下に、正記憶装置と副記憶装置が混在することもあるが、この場合には、その制御装置１１０は、正記憶装置に対して正制御装置であり、副記憶装置に対して副制御装置となる。

【００１８】制御装置１１０は、中央処理装置１００および他の制御装置１１０とリンク１５０で接続されている。リンク管理機構１２０は、使用すべきリンク１５０の切替え等を行なう。キャッシュ１７０は、中央処理装置１００からライトされたデータあるいは記憶装置１３０からリードされたデータを一時的に格納しておくメモリである。中央処理装置１００からアクセスのあったデータが、キャッシュ１７０上に存在していれば記憶装置１３０へのアクセスは行なわない。プロセッサ１６０は、記憶装置１３０，キャッシュ１７０，中央処理装置１００，他の制御装置１１０間のデータ転送を制御する。ディレクトリ１８０は、キャッシュ１７０の管理情報を格納するためのメモリである。プロセッサ１６０は、ディレクトリ１８０をアクセスすることによって、ヒットミス判定等を行なう。

【００１９】ヒットミス判定とは、中央処理装置１００からアクセス要求のあったデータの位置すなわち記憶装置アドレス，データアドレス，データ長から、目的とするデータがキャッシュ１７０上に存在する（ヒット）か否か（ミス）を判定する処理を言う。

【００２０】記憶装置１３０は、磁気ディスク装置である。但し、磁気テープ装置や光ディスク装置などでもかまわない。図２は、磁気ディスク装置の説明図である。この磁気ディスク装置２５０は、データを記録する複数個のディスク２４０と、ディスク２４０上のデータをリード・ライトするためのヘッド２３０と、制御装置インタフェース２２０から構成される。ディスク２４０が１回転する間にヘッド２３０がアクセス可能な円状の記録単位をトラック２００と呼ぶ。トラック２００はディスク２４０上に複数個存在する。各トラックはセクタ２６０と呼ばれるある決まった大きさの領域に分割されていて、そのセクタがデータアクセスの最小単位（スロット）となる。あるトラック２００がアクセス対象になった時、そのトラック２００をリードライトできる位置にヘッド２３０を移動する。この動作をシークと言う。このディスク装置２５０では、複数のヘッド２３０が同時に移動する。ヘッド２３０は上から昇順に番号付けされていて、それをヘッド番号と呼ぶ。ディスク２４０が１回転する間にすべてのヘッド２３０がアクセス可能なトラック２００の集まりをシリンダ２１０と呼ぶ。シリンダ２１０にはディスク２４０の内側から昇順に番号付けされていて、それをシリンダ番号と呼ぶ。

【００２１】ディレクトリ１８０内には、ヒットミス判定テーブルが格納されている。図３は、ヒットミス判定テーブルの構造例である。このヒットミス判定テーブル３４０は次の表から構成されている。

装置アドレス対応表３００：ディスク装置２５０毎にエントリを持ち、それぞれのエントリは当該ディスク装置に対応するシリンダ番号対応表３１０へのポインタを格納している。

シリンダ番号対応表３１０：一つのディスク装置のシリンダ毎にエントリを持ち、それぞれのエントリは当該シリンダに対応するトラック番号対応表３２０へのポインタを格納している。

トラック番号対応表３２０：一つのシリンダのトラック毎にエントリを持ち、それぞれのエントリは当該トラックに対応するスロット管理表３３０へのポインタを格納している。

スロット管理表３３０：一つのトラックのキャッシュ管理情報であり、次の情報をもつ。

データ有無マップ３３１：トラック上のどのセクタがキャッシュ１７０上に存在するかを示す情報である。

キャッシュアドレス３３２：トラック上のセクタがキャッシュ１７０上に存在するとき、そのデータが格納されているアドレス情報である。

ポインタ３３３：キュー管理等を行なう時に使用する情報である。

時刻３３４：当該スロットに対して正記憶装置１３０ａにおいてライトのあった時刻を副制御装置側１１０ｂ側で管理するための情報である。

RVOLダーティ状態３３５：当該スロットが副記憶装置１３０ｂに対して未更新かどうかを示す情報である。

トラックアドレス３３６：当該トラックのアドレスである。

ヘッドアドレス３３７：当該トラックにアクセスするヘッドのアドレスである。

【００２２】中央処理装置１００から指定された位置付け情報すなわちアクセス対象の記憶装置アドレス，シリンダ番号，トラック番号を基に、装置アドレス対応表３００，シリンダ番号対応表３１０，トラック番号対応表３２０，スロット管理表３３０をたどり、データ有無マップ３３１の対応するビットがオンであればヒットと判定し、オフであればミスと判定する。さらに、データ転送を行なう場合には、キャッシュアドレス３３２に目的のセクタに対応するオフセットを加えることによって、転送対象となるキャッシュアドレスを算出する。

【００２３】非同期コピーによる遠隔二重書きを実現するために、副記憶装置１３０ｂへの未更新のデータ（これをRVOLダーティデータという）を管理するRVOLダーティ管理表をディレクトリ１８０にもつ。

【００２４】図４は、RVOLダーティ管理表の構成図であ

る。このRVOLダーティ管理表４００は、記憶装置VOL1, VOL2, …毎のエントリを有している。一つのエントリには、RVOL情報４０１と、RVOLダーティリストヘッド４０２とが格納されている。RVOL情報４０１には、当該記憶装置を正としたときの副制御装置１１０ｂアドレスおよび副記憶装置１３０ｂアドレスを保持する。当該記憶装置１３０が正／副のペアを組んでいない場合は特殊な値が格納される。RVOLダーティリストヘッド４０２は、未更新のスロットを管理するためのRVOLダーティリスト４１０を指す。未更新のデータが存在しない場合にはＮＵＬＬ値が格納される。RVOLダーティリスト４１０は、スロット管理表３３０内のポインタ３３３を用いて未更新スロットのスロット管理表３３０をリスト構造にしたものである。

【００２５】RVOLダーティ管理表４００の登録は、中央処理装置１００から正記憶装置１３０ａへのライト系アクセス時もしくはライト系アクセスと判明した時に行なわれる。中央処理装置１００は、定められたプロトコルに従って、正記憶装置１３０ａへアクセスする。そのプロトコルについては様々なものが存在するが、例えばメインフレームを中心とした大型計算機システムでは、ＣＫＤプロトコルが一般的に使用されている。このＣＫＤプロトコルでは、データアクセスコマンドに先だって、位置付け情報の他に、アクセス形態（リードまたはライト系アクセス）を表す情報を含む位置付けコマンドが発行される。従って、ライト系アクセスかどうかは、位置付けコマンドを受領した時点で判明するため、この時にRVOLダーティ管理表４００の登録を行なう。また、ワークステーションやパソコンでは、データアクセスコマンドに位置付け情報を含むＦＢＡプロトコルが用いられている。このＦＢＡプロトコルの場合には、データアクセスコマンド受領後、RVOLダーティ管理表４００の登録を行なう。

【００２６】図５は、RVOLダーティ管理表４００の登録処理のフローチャートである。ステップ５１０では、上述のようにライト系アクセスか否かを判定し、ライト系アクセスでなければ処理を終了し、ライト系アクセスであることが判明すればステップ５２０へ進む。ステップ５２０では、アクセス対象の記憶装置１３０のアドレスからRVOLダーティ管理表４００のエントリを求め、当該エントリのRVOL情報４０１を参照することによって、アクセス対象の記憶装置１３０が正／副のペアを形成しているかどうかをチェックする。ペアを形成していなければ処理を終了し、ペアを形成していればステップ５３０へ進む。ステップ５３０では、RVOL情報４０１から副制御装置１１０ｂのアドレスを求めて、その副制御装置１１０ｂに対して位置付け情報を転送する。これにより、副記憶装置１３０ｂに未反映のデータが存在することを副制御装置１１０ｂに通知することが出来る。ステップ５４０では、RVOLダーティ管理表４００にライト対象スロットを登録する。すなわち、RVOLダーティリストヘッダ４０２の指すスロット管理表３３０の前に、新規のスロット管理表３３０を挿入する。

【００２７】なお、前記ステップ５３０において、副制御装置１１０ｂに位置付け情報を転送する処理は、見かけ上のオーバヘッドを削減するために、中央処理装置からライトデータを受けとる処理と並行して行なうのが良い。例えばＣＫＤプロトコルで説明すると、位置付けコマンドにより位置付け情報を受領した後、子ジョブを生成して位置付け情報転送処理を実行させておき、自ジョブでは位置付けコマンドにチェインするライトコマンドの処理を実行する。自ジョブでは、当該コマンドチェインが完了した後、子ジョブの完了を待ってから中央処理装置１００に対してライト処理完了を報告する。

【００２８】また、前記ステップ５３０において、副制御装置１１０ｂに対して、位置付け情報の他に、現在の時刻を転送してもよい。これにより、どの時点でデータが消失したかを副制御装置１１０ｂ側で後に調べることが出来るようになる。さらに、現在の時刻を転送する際、正制御装置１１０ａの内部時間を転送してもよいが、外部時間との誤差をなくすために、例えば中央処理装置１００からの位置付けコマンドパラメータ内に時刻を格納しておき、その時刻を転送してもよい。例えば図６に示すRVOLダーティ情報６００により、位置付け情報および時刻は、正制御装置１１０ａから副制御装置１１０ｂへ転送される。このRVOLダーティ情報６００は、RVOLダーティ情報の転送であることを表すコマンドコード６０１と、副記憶装置アドレス６１０と、ダーティデータ位置情報６２０と、時刻６３０とから構成される。

【００２９】図７は、RVOLダーティ管理表４００の削除処理のフローチャートである。なお、効率的な非同期コピー方式の遠隔二重書きを実現するために、実装している正記憶装置１３０ａ対応にRVOLライトを専属に実行する非同期二重書きジョブが存在するものとする。この非同期二重書きジョブは、一定周期毎に、ランとスリープとを繰り返すジョブで、ラン中にRVOLダーティ管理表４００の削除処理を実行する。ステップ７１０では、実行中の非同期二重書きジョブに対応する正記憶装置１３０ａのRVOLダーティ管理表４００のRVOLダーティリストヘッダ４０２を参照し、その値がＮＵＬＬならばRVOLダーティデータは存在しないので、処理を終了する。RVOLダーティリストヘッダ４０２の値がＮＵＬＬでないならば、RVOLダーティデータが存在するので、ステップ７２０へ進む。ステップ７２０では、RVOLダーティ管理表４００のRVOL情報４０１と当該スロット管理表３３０の情報とから副記憶装置１３０ｂに対するライトコマンドを生成し、RVOLライト処理を実行する。なお、ライトコマンドの生成方法については例えばＵＳＰ５５５８４５に記載されている。ステップ７３０では、ライト完了したスロット管理表３３０をRVOLダーティリスト４１０から

削除する。処理が完了すると、しばらくスリープする。

【００３０】以上が正制御装置１１０ａ側の処理である。次に、副制御装置１１０ｂ側の処理を説明する。

【００３１】副制御装置１１０ｂの主な処理は、RVOLダーティ情報６００が正制御装置１１０ａから転送されてきてから、当該データがライトされるまでの間、RVOLダーティ情報６００を保持しておくことである。この処理を実現するため、副制御装置１１０ｂでも、図４に示したRVOLダーティデータ管理表４００と基本的に同一のテーブル構造をもつ。RVOLダーティリスト４１０に接続されたスロット管理表３３０が、副記憶装置１３０ｂへ未反映のスロットを表す。なお、副制御装置１１０ｂでは、RVOL情報４０１は、使用されない。

【００３２】RVOLダーティデータ管理表４００への登録は、正制御装置１１０ａからRVOLダーティ情報６００が転送されてきた時に行なわれる。すなわち、RVOLダーティデータ情報６００の転送かどうかをコマンドコード６０１によって判定し、RVOLダーティ情報６００の転送であれば、その記憶装置アドレス６１０から、対応する副記憶装置１３０ｂのRVOLダーティ管理表４００およびRVOLダーティリスト４１０を求める。また、ダーティデータ位置情報６２０からヒットミス判定を実行すし、対象となるスロット管理表３３０を得る。そして、そのスロット管理表３３０を前記RVOLダーティリスト４１０に登録すると共に、そのスロット管理表３３０のRVOLダーティ状態３３５を“ダーティあり”とする。

【００３３】スロット管理表３３０の削除は、正制御装置１１０ａからの対象スロットに対するライトを実行した時に行なわれる。すなわち、正制御装置１１０ａからのライト要求に対して、位置付け情報を基にヒットミス判定を行ない、得られたスロット管理表３３０のRVOLダーティ状態３３５を参照し、“ダーティあり”ならば、ライト処理を実行した後、RVOLダーティリスト４１０から当該スロット管理表３３０を削除し、RVOLダーティ情報３３５を“ダーティなし”に変更する。

【００３４】以上によって、副記憶装置１３０ｂに対する未更新データの位置情報およびそのデータが未更新となった時刻が、正制御装置１１０ａおよび副制御装置１１０ｂでそれぞれ保持されることになる。

【００３５】図８に、消失データの位置情報，消失時刻を記録するための消失データリストを示す。この消失データリスト１１００は、記憶装置アドレス１１１０と、消失データ数１１２０と、消失データ毎のデータアドレス１１３０および消失した時刻１１４０とから構成される。

【００３６】なお、前記消失データ数が所定の閾値を越えると、記憶装置１３０全体が無効になり、記憶装置１３０全体を回復するようにしてもよい。こうすると、消失データリスト１１００が膨大になるのを防ぐことが出来る。前記閾値は、保守員などから設定できるのが好ましい。

【００３７】副制御装置１１０ｂは、中央処理装置１００からの消失データリスト１１００の読み出し要求に対して、消失データリスト１１００を次の手順により作成する。

（１）要求のあった記憶装置アドレスから、対応するRVOLダーティ管理表４００を検索する。

（２）検索したRVOLダーティ管理表４００のRVOLダーティリストヘッド４０２の値がＮＵＬＬなら、消失データは存在しないので、消失データ数１１２０に“０”を記録する。

（３）RVOLダーティリストヘッド４０２の値がRVOLダーティリスト４１０を指しているなら、そのRVOLダーティリスト４１０に接続しているスロット管理表３３０のトラックアドレス３３６，ヘッドアドレス３３７をデータアドレス１１３０に格納し、時刻３３４を時刻１１４０に格納する。これを、RVOLダーティリスト４１０をたどりながら、繰り返し、消失データ数をカウントしていく。

（４）カウントしていた消失データ数を消失データ数１１２０に格納する。

（５）指定された記憶装置アドレスを記憶装置アドレス１１１０に格納する。

【００３８】以上により消失データリスト１１００が完成すると、副制御装置１１０ｂは、中央処理装置１００に消失データリスト１１００を返送する。そこで、保守員は、消失データリスト１１００を読み出すことにより、消失データの有無を調べることが出来る。そして、データが消失していた場合には、データアドレス１１３０からデータを回復することが出来る。また、時刻１１４０の最も古いものを検索して、その世代のバックアップファイルからデータを回復することも可能となる。

【００３９】図９は、正制御装置がアクセス不可能になっている場合の副制御装置１１０ｂにおけるアクセス可否判定処理８００を表すフローチャートである。ステップ８１０では、アクセス対象スロットのヒットミス判定を実行して、スロット管理表３３０を得る。ステップ８２０では、前記スロット管理表３３０のRVOLダーティ状態３３５から、当該スロットのデータが副記憶装置１３０ｂに未反映かどうかを判定する。未反映であれば、データが消失しているため、ステップ８３０へ進む。未反映でなければ、データが消失していないため、ステップ８４０へ進む。

【００４０】ステップ８３０では、中央処理装置１００に対して異常終了を報告する。ステップ８４０では、アクセスを許可する。

【００４１】以上で第１の実施形態の説明を終るが、ここで第１の実施形態の要点をまとめると次のようになる。

（１）正制御装置１１０ａからライト処理を実際に行う

前に、副制御装置１１０ｂへ位置付け情報および時刻を転送しておく。副制御装置１１０ｂは、実際にライトが行なわれるまで、前記位置付け情報と前記時刻を保持しておく。

（２）障害、災害等により正記憶装置１３０ａがアクセス不能となった場合は、副制御装置１１０ｂから消失データの有無を判断する。消失データが有った場合には、その消失データの時刻から、回復すべきバックアップファイルを決定する。

（３）消失データに対してアクセスがあった場合は、アクセスを許可せず、異常終了を報告する。

【００４２】従来の非同期コピー方式の遠隔二重書きでは、消失したデータの有無を判断できないため、消失データが存在するにもかかわらずデータの回復を行なわなかった場合には、データ化けが発生する。一方、消失データが存在しないにもかかわらずデータの回復を行なった場合には、回復作業が無駄になり、データが数世代前に戻るため、極めてコストがかかる。また、記憶装置１３０全体を回復しなくてはならず、効率が悪い。これに対して、上記第１の実施形態による非同期コピー方式の遠隔二重書きでは、副側のサイトで運用を開始する前に、まず正記憶装置１３０ａと副記憶装置１３０ｂの内容が一致しているかどうかをチェックして、不一致であれば適切な時刻のバックアップファイルからデータ回復してから運用を開始する。従って、前記データ化けや無駄なデータ回復を回避することが出来る。以上により、災害や障害等によって正記憶装置１１０ａに対するアクセスが不可能になった場合でも、副記憶装置１３０ｂを使用して業務を副側のサイトで引き継ぐことができ、しかも、システム停止時間を最小にできる。

【００４３】－第２の実施形態－

前記第１の実施形態では、中央処理装置１００からの１回のライト要求毎（ＣＣチェイン毎）に副制御装置１１０ｂに位置付け情報を転送していたため、正制御装置１１０ａと副制御装置１１０ｂの間の通信回数が多くなり、性能が低下する。これを解決するため、第２の実施形態では、正記憶装置１３０ａと副記憶装置１３０ｂの内容が一致あるいは不一致のときのみ、そのことを副制御装置１１０ｂに通知し、正制御装置１１０ａと副制御装置１１０ｂの間の通信回数を削減する。

【００４４】記憶装置１３０は、二重書き一致または二重書き不一致の２状態をとる。ここで、二重書き一致とは、正記憶装置１３０ａと副記憶装置１３０ｂの内容が一致した状態であることを示す。また、二重書き不一致とは、正記憶装置１３０ａと副記憶装置１３０ｂの内容が不一致の状態であることを示す。これらの二重書き状態は、正制御装置１３０ａおよび副制御装置１３０ｂのRVOLダーティ管理表４００のRVOL情報４０１に格納しておく。また、二重書き不一致となった時刻も、RVOL情報４０１に格納しておく。

【００４５】図１０は、二重書き一致から二重書き不一致に状態遷移させる不一致化状態変更処理９００のフローチャートである。ステップ９１０では、ライト系アクセスかどうかチェックする。ライト系アクセス以外であれば、二重書き状態の変更は発生しないので、処理を終了する。ライト系アクセスであれば、ステップ９２０へ進む。ステップ９２０では、アクセス対象の記憶装置１３０が正／副のペアを形成しているかどうかをRVOL情報４０１から判定する。ペアを形成していなければ、二重書き状態の変更は発生しないので、処理を終了する。ペアを形成していた場合は、ステップ９３０へ進む。ステップ９３０では、現在の二重書き状態が二重書き一致かどうかをRVOL情報４０１から判断する。二重書き不一致であれば、二重書き状態の変更は発生しないので、処理を終了する。二重書き一致であれば、ステップ９４０へ進む。ステップ９４０では、副制御装置１１０ｂに対して、二重書き不一致となったことを通知する。効率の観点から、この通知の処理と、中央処理装置１００からのライト処理とを、並行して行なうのが好ましい。ステップ９５０では、RVOL情報４０１を二重書き不一致に状態変更する。

【００４６】なお、上記不一致化状態変更処理９００で、二重書き不一致の通知だけでなく、その時刻を副制御装置１１０ｂに通知してもよい。この場合、中央処理装置１００からライト系アクセスを受けた時刻を副制御装置１１０ｂへ転送し、かつRVOL情報４０１にも当該時刻を記録する。なお、この時刻は、正制御装置１１０ａの内部時刻でも良いし、中央処理装置１００から与えられる時刻でも良い。

【００４７】図１１は、二重書き不一致から二重書き一致に状態遷移させる一致化状態変更処理１０００のフローチャートである。この一致化状態変更処理１０００は、第１の実施形態で説明した非同期二重書きジョブが副記憶装置１３０ｂに未更新データをライトした時に実行される。ステップ１０００では、第１の実施形態で説明したように未更新データを副記憶装置１３０ｂに対してライトする。ステップ１０１０では、第１の実施形態で説明したようにライト完了したスロット管理表３３０をRVOLダーティリスト４１０から外す。ステップ１０２０では、RVOLダーティリスト４１０をたどることにより、副記憶装置１３０ｂに対する未更新データが存在するかどうかを判断する。RVOLダーティリスト４１０にスロット管理表３３０が未だ接続されていれば、二重書き不一致のままであるので、処理を終了する。RVOLダーティリスト４１０にスロット管理表３３０が接続されていなければ、二重書き不一致が解消されたので、ステップ１０３０へ進む。ステップ１０３０では、二重書き一致となったことを副制御装置１１０ｂに通知する。ステップ１０４０では、RVOL情報４０１を二重書き一致に状態変更する。

【００４８】図１２は、正制御装置１１０ａから副制御装置１１０ｂに対する状態変更の通知で使用される状態変更通知情報の例示図である。この状態変更通知情報１３００は、状態変更通知コマンドであることを示すコマンドコード１３０１と、対象となる副記憶装置アドレス１３１０と、状態変更を行なった時刻１３２０と、新状態のコードを表す状態コード１３３０とが格納される。

【００４９】以上が正制御装置１１０ａ側の処理である。次に、副制御装置１１０ｂ側の処理を説明する。

【００５０】副制御装置１１０ｂでは、状態変更通知情報１３００を受領すると、対象となる副記憶装置１３０ｂに対応したRVOL情報４０１を状態コード１３３０で指定された状態に変更する。なお、副制御装置１１０ｂでは、RVOLダーティリスト４１０は使用しない。

【００５１】図１３に、記憶装置１３０の二重書き状態を中央処理装置１００に通知するための正副状態情報１２００を示す。この正副状態情報１２００は、記憶装置アドレス１２１０と、二重書き一致あるいは二重書き不一致を表す二重書き状態１２２０と、二重書き不一致となった時刻を表す状態変更時刻１２３０とから構成される。

【００５２】副制御装置１１０ｂは、中央処理装置１００からの正副状態情報の読み出し要求に対して、正副状態情報１２００を次の手順により作成する。

（１）指定された記憶装置アドレスから、対応するRVOLダーティ管理表４００を検索し、RVOL情報４０１から、二重書き状態を得て、二重書き状態１２２０へ格納する。

（２）二重書き不一致であれば、RVOL情報４０１に記録された時刻を状態変更時刻１２３０へ格納する。

以上により正副状態情報１２００が完成すると、副制御装置１１０ｂは、中央処理装置１００に正副状態情報１２００を返送する。そこで、保守員は、正副状態情報１２００を読み出すことにより、二重書き状態を確認することが出来る。そして、二重書き不一致であれば、不一致となった時刻から最も世代の近いバックアップファイルを選び出し、データ回復を行なうことが出来る。この第２の実施形態の方法は、個々のデータ回復ができないかあるいは不要な場合に有用である。

【００５３】中央処理装置１００から二重書き不一致の副記憶装置１３０ｂに対してアクセス要求があったら、その副制御装置１１０ｂは、要求のあった記憶装置アドレスから対応するRVOLダーティ管理表４００を求め、そのRVOL情報４０１から、当該記憶装置１３０が正／副のペアを組んでいるかどうか、および、正／副のペアを組んでいる場合には二重書き不一致かどうかを調べる。そして、二重書き不一致の場合は、旧世代のデータを転送するのを防ぐために、アクセスを拒否する。

【００５４】以上で第２の実施形態の説明を終るが、ここで第２の実施形態の要点をまとめると次のようになる。

（１）正および副記憶装置１３０の内容の一致／不一致を表す二重書き状態を制御装置１１０に保持する。

（２）二重書き状態は、副記憶装置１３０ｂに未更新データができる時に二重書き不一致となり、未更新データがなくなった時に二重書き一致となる。

（３）障害や災害により正記憶装置１３０ａへのアクセスが不可能となった場合は、副制御装置１１０ｂから二重書き状態を読み出し、二重書き不一致となった時刻から最も近い世代のバックアップファイルからデータを回復することが出来る。

（４）二重書き不一致となっている副記憶装置１３０ｂに対する中央処理装置１００からのアクセスは拒否される。

【００５５】上記第２の実施形態では、副側のサイトで運用を開始する前に、保守員は、正記憶装置１３０ａと副記憶装置１３０ｂの内容が一致しているか不一致なのかをチェックする。そして、不一致であれば、適切な時期のバックアップファイルからデータを回復してから、運用を開始する。以上により、災害や障害等によって正記憶装置１１０ａに対するアクセスが不可能になった場合でも、副記憶装置１３０ｂを使用して業務を副側のサイトで引き継ぐことができ、しかも、システム停止時間を最小にできる。

【００５６】－第１の実施形態と第２の実施形態の差異－

（１）第１の実施形態では、中央処理装置１００からのライト処理の応答時間が劣化する可能性がある。一方、第２の実施形態では、応答性能劣化はほとんどない。

（２）障害、災害後の消失データの回復では、第１の実施形態では消失したデータ単位に回復可能なのに対して、第２の実施形態では、記憶装置全体を回復する必要がある。

【００５７】

【発明の効果】本発明の計算機システムによれば、副記憶装置に対する未更新データができると、その位置付け情報あるいは正副の二重書き状態が不一致であることを、正制御装置から副制御装置に通知するので、障害や災害等により正記憶装置が使用不能となった時に、消失データの有無を副制御装置から判断でき、適切に対処することが可能になる。

【図面の簡単な説明】

【図１】本発明の第１の実施形態にかかる計算機システムの構成図である。

【図２】磁気ディスク装置の構成図である。

【図３】ヒットミス判定テーブルの構成図である。

【図４】RVOLダーティ管理表の構成図である。

【図５】RVOLダーティ管理表の登録処理を表すフローチャートである。

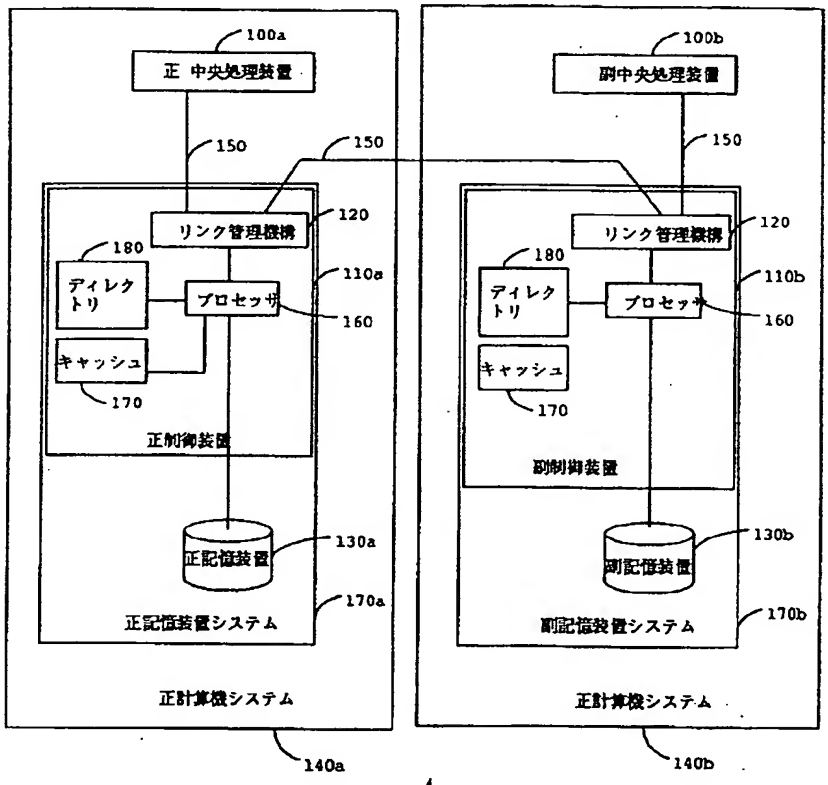
【図６】RVOLダーティ情報の構成図である。

17　　18

【図７】RVOLダーティ管理表の削除処理を表すフローチャートである。
【図８】消失データリストの構成図である。
【図９】アクセス可否判定処理を表すフローチャートである。
【図１０】不一致化状態変更処理を表すフローチャートである。
【図１１】一致化状態変更処理を表すフローチャートである。
＊【図１２】状態変更通知情報の構成図である。
【図１３】正副状態情報の構成図である。
【符号の説明】
１００…中央処理装置
１１０…制御装置
１３０…記憶装置
１６０…プロセッサ
１７０…キャッシュ
＊　１８０…ディレクトリ

【図１】

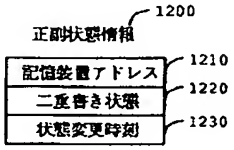
図1



1
計算機システム

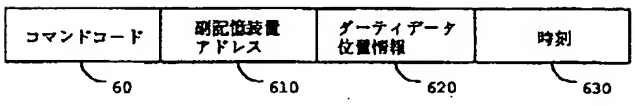
【図１３】

図13 正副状態情報



【図６】

図6



RVOL ダーティ情報　600

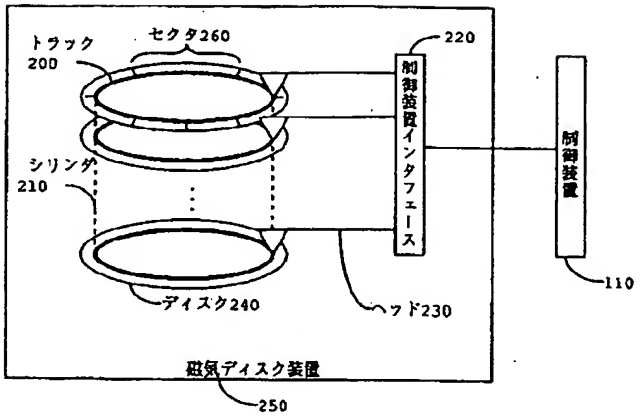
【図１２】

図12

状態変更通知情報
1300

【図2】

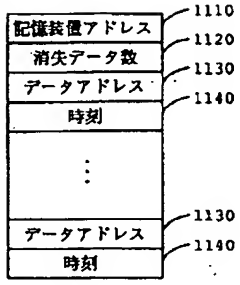
図2



【図8】

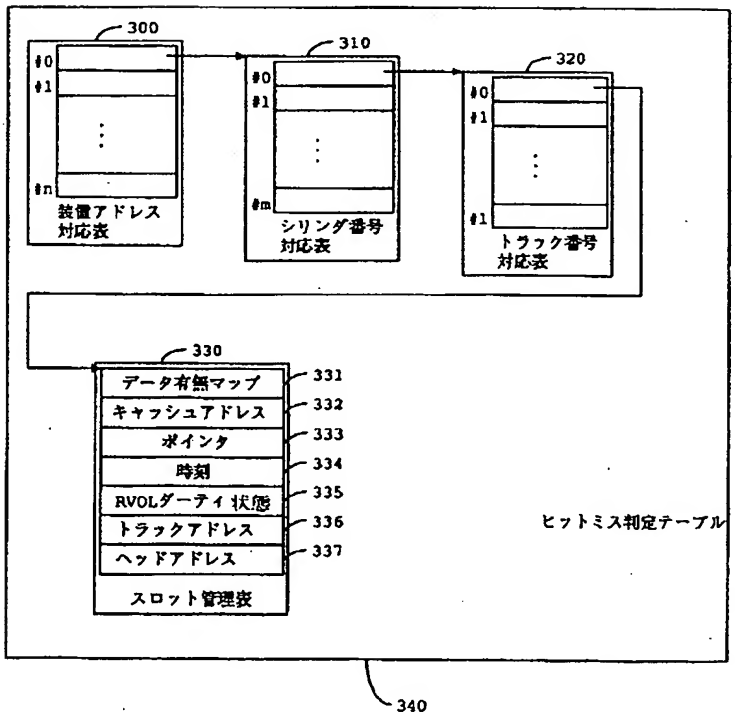
図8

消失データリスト
1100



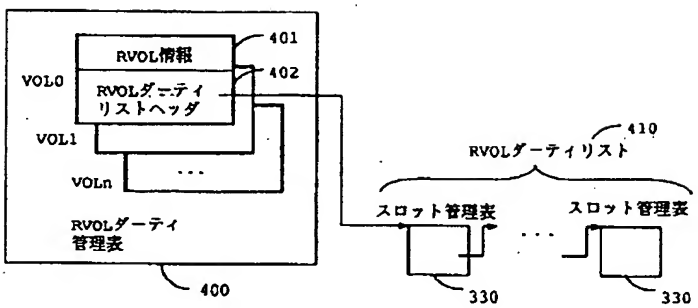
【図3】

図3



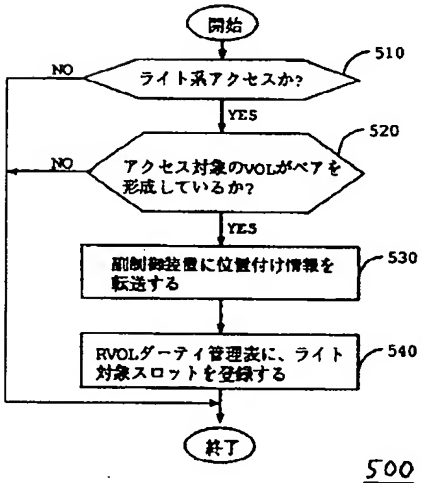
【図4】

図4



【図5】

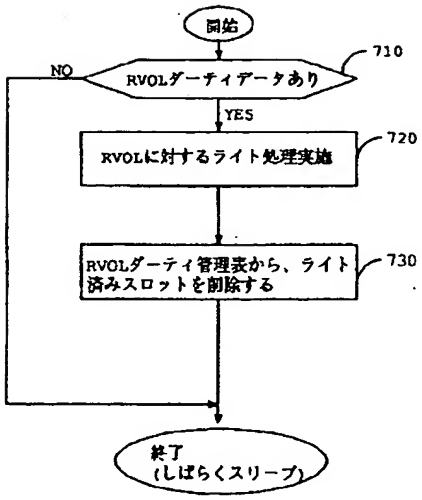
図5



RVOLダーティ管理表登録処理

【図7】

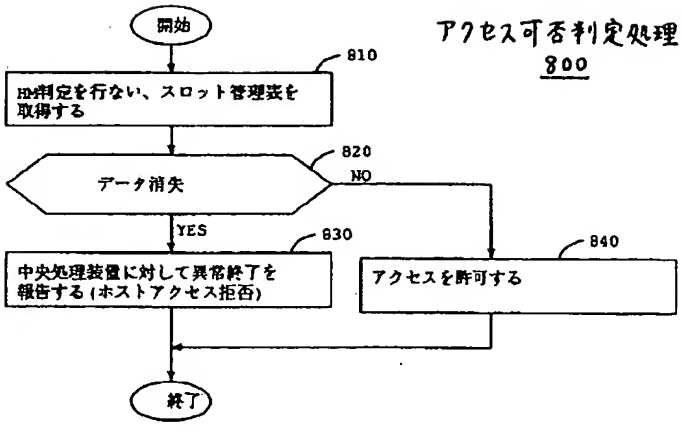
図7



RVOLダーティ管理表削除処理
700

【図9】

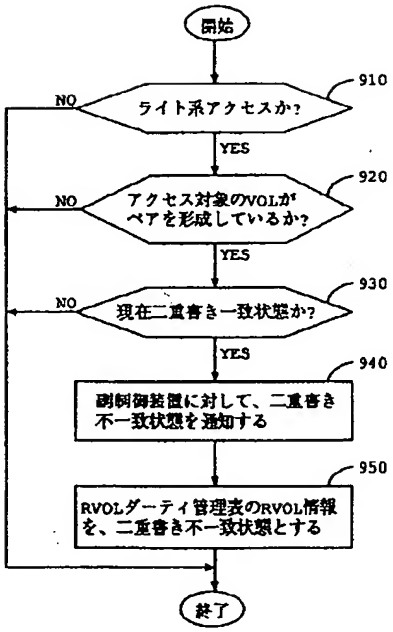
図9



アクセス可否判定処理
800

【図１０】

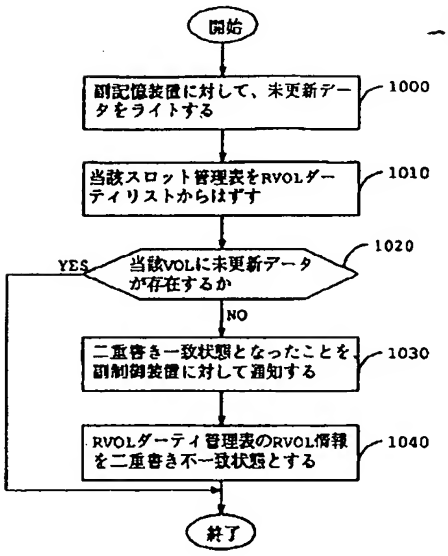
図10



不一致化状態変更処理
900

【図１１】

図11



一致化状態変更処理